基于类别信息和特征熵的文本特征权重计算 *

阿力木江·艾沙 [a, b]，殷晓雨 [b]，库尔班·吾布力 [b]，李　喆 [a]

(新疆大学 a. 网络与信息技术中心; b. 信息科学与工程学院, 乌鲁木齐 830046)

摘　要：文本向量化是文本分类的基础，特征权重是直接影响文本向量表示质量的重要因素之一。基于类别信息的特征权重计算方法对特征与类别的关系表达不够准确，即对于类别频率相同的特征无法比较其对类别的区分能力，因此要考虑特征在类内的分布情况。将特征的反类别频率（inverse category frequency，ICF）和类内熵（entropy）相结合引入到特征权重计算方案中，构造了两种有监督特征权重计算方案。在维吾尔文文本分类语料上进行的实验结果表明，该方法能够明显改善样本的空间分布状态并提高维吾尔文文本分类的微平均 F1 值。

关键词：文本分类；文本特征；权重计算；类别频率

中图分类号：TP391.1　　doi: 10.3969/j.issn.1001-3695.2018.05.0294

Feature weighting scheme based on category information and term entropy

Alimjan Aysa[a, b], Yin Xiaoyu[b], Kurban Ubul[b], Li Zhe[a]

(*a. Network & Information Technology Center*, *b. School of Information Science & Engineering Xinjiang University*, *Urumqi* 830046, *China*)

Abstract: Text vectorization is the basis of text classification. Feature weighting is one of the important factors that directly affect the quality of text vector representation. Feature weighting schemes based on category information is not accurate enough to express the relationship between features and categories. That is the classification ability of the features with the same category frequency can't be compared, so the distribution of the features in the category should be considered. This paper combines the inverse category frequency (ICF) and inner category entropy of the features into the term weight calculation, and constructs two supervised feature weighting schemes. The experimental results on the Uygur text categorization dataset showed that this method can obviously improve the spatial distribution of the samples and improve the micro average F1 value of the Uygur text classification.

Key words: text classification; text feature; term weighting; category frequency

0　引言

在文本分类中，首先将自然语言文本转换成一种计算机能够处理的内部表示形式，分类器才能“理解”文本内容并进行分类操作。这个过程叫做文本向量化或文本表示。目前，向量空间模型（vector spase model，VSM）仍然是主流的文本表示方法。在 VSM 中，文本被表示为在特征空间 $d=\{w_1,w_2,..,w_n\}$ 中的一个向量，其中 n 为特征集的大小。特征权重 w_i 表示特征 t_i 在文本 d 中的重要程度。特征权重计算方法经常影响分类器的效率。文献[1]指出特征权重方案的选择很大程度上影响分类器分类精度。因为，一个好的特征权重方案会给每一个被选的特征项指定一个合理的权重。

文本特征的权重综合反映了该特征对标识文本内容的贡献度和区分文本类别的能力[2]。合理的特征权重计算方法能使特征对分类的作用变得更加显著[3]。因此，特征权重计算对文本分类起着至关重要的作用。特征权重计算是文本分类领域的研究热点之一[4,5]。

1　相关研究工作

文本分类中的权重计算方案来自于信息检索（information retrieval，IR）领域。其中最有名的 *tf.idf*（term frequency and inverse document frequency）算法在 IR 中获得了很大的成功。正因为 *tf.idf* 在 IR 中的成功，研究人员将 *tf.idf* 原封不动的在文本分类任务中使用。而且在很多文本分类任务中，把 *tf.idf* 作为默认的权重方案在使用。也有一些新的改进方案被提出。最早，Debole 和 Sebastiani[6]针对文本分类任务提出了一种构造有监

收稿日期：2018-05-07；**修回日期：**2018-06-26　　**基金项目：**新疆维吾尔自治区自然科学基金资助项目（2016D01C068）

作者简介：阿力木江·艾沙（1973-），男，维吾尔族，教授，博士，硕导，主要研究方向为自然语言处理、信息安全（alim@xju.edu.cn）；殷晓雨（1992-），男，回族，硕士研究生，主要研究方向为文本挖掘；库尔班·吾布力（1974-），男，维吾尔族，教授，硕导，主要研究方向为文本图像处理；李喆（1978-），女，讲师，主要研究方向为计算机应用技术.

督特征权重计算方案的方法，就是用特征选择函数 χ^2 、信息增益（information gain, IG）和增益率（gain ratio，GR）来替换 *tf.idf* 算法中的 *idf* 项。文献[7]将 *tf.idf* 和 IG 相结合，改进了 *tf.idf* 方案。Lan 等人[8]提出了 *tf.rf*（term frequency and relevance frequency）并改进了英文文本分类的表现。该方案只考虑了相关文本，而忽略特征在非相关文本中的分布情况。不过，它们在英文标准语料库上实验结果显示，*tf.rf* 方法表现出比其他有监督特征权重方案（如：*tf.logOR*、*tf.* χ^2 、*tf.ig*）和传统方案（如：*tf.idf*、*tf*）更好地性能。除此之外，文献[9]提出了一种基于概率的有监督特征权重方案叫做 *prob* 方案并改进了针对非平衡数据集的文本分类性能。文献[10]针对问题分类提出了三种新的有监督权重方案并在方案中用到了 *icf* 因子。文献[11]提出了基于逆类空间密度频率（inverse class space density frequency，ICSDF）的两个新的特征权重计算方法 *tf* ICSDF* 和 *ICSDF-based*。该方法相比传统的特征加权方法(*prob-based*、*tf.icf* 和 *icf-based*)能够有效地提升文本分类性能。文献[12]提出了平均反类别频率的概念考虑了特征在不同词频下的局部类别频率。但反类别频率算法只关注特征是否在类别内出现过，并不考虑特征在该类别内出现的文本中的分布情况，夸大了类内低频文档对分类的作用，这是大部分引入 *icf* 因子算法的局限性。

本文对维吾尔文文本分类中的特征权重计算问题进行了研究。针对已有的基于 *icf* 的特征权重计算方法的局限性，本文将特征的反类别频率（inverse category frequancy，*icf*）和类内熵引入到已有的特征权重计算方案中，构造了两种有监督特征权重计算方案。采用本文方案和已有的权重方案在维吾尔文文本分类数据集上进行实验测试，并分析了实验结果。

2　引入类别信息和熵的特征权重计算方案

将反类别频率（inverse category frequency，*icf*）和特征熵引入到文本分类中的特征权重计算中。

Category frequency(*cf*)：类别频率是指特征 t_i 出现的类别数；

Inverse category frequency(*icf*)：反类别频率 *icf* 的计算公式和 *idf* 类似，如下公式所示：

$$icf(t_i) = lb(\frac{|C|}{cf(t_i)})$$

其中，|C|为训练集中的类别总数。

在特征权重计算中引入 *icf* 是基于这样的一个假设：特征 t_i 出现的类别数越少，它携带的类别信息量就越大。这个假设被叫做 *icf* 假设，其在类别级上侧重于中低频特征而仰止高频特征。但是，*icf* 因子只考虑特征在类间的分布情况，但并没有考虑特征在每一类内的分布情况。比如，一个特征集中出现在某个类别中的极少数几个文本中，那么这个特征不能很好的代表这个类，应该给它分配较小的权重。所以还要考虑特征的类内分布情况。

若某一个特征词在类内分布越均匀，则该词越能代表该类，具有较高的类别区分能力，它在该类别的所有文档中都应当被赋给更大的权重。相反，如果某个特征词仅在类内少数几个文本中集中出现，则该特征词不能很好地代表该类别，特征词类别区分能力低，应该分配较低的权重。通过分析可知，类内特征词分布熵值的大小与特征词所能提供的分类信息量是一致的[13]，即熵值越大，特征词所提供的类别信息量越大，越能代表该类。

特征 t_i 在类 c_k 中的熵 *te*（*term entropy*）定义为

$$te(t_i,c_k) = -\sum_{j=1}^{|c_k|}\frac{tf(t_i,d_{kj})}{tf(t_i,c_k)}lb\frac{tf(t_i,d_{kj})}{tf(t_i,c_k)}$$

$$tf(t_i,c_k) = \sum_{j=1}^{|c_k|}tf(t_i,d_{kj})$$

其中: $tf(t_i,d_{kj})$ 为特征 t_i 在 c_k 类中的第 j 个文本中出现的频数，$|c_k|$ 表示 c_k 类中的文本总数，$tf(t_i,c_k)$ 表示特征 t_i 在 c_k 类文本中出现的总频数。

可以看出，当特征词 t_i 在类别 c_k 中的每个文本中都出现时，$te(t_i,c_k)$ 值取最大值，分类能力最强；当特征词仅在类别中的某一个文本中集中出现时，$te(t_i,c_k)$ 值取最小值，分类能力最弱。因此，$te(t_i,c_k)$ 很好地反映了特征词在类内的分布情况，并且其值与特征的类别区分能力呈正比。

根据以上的分析，我们将 *icf* 因子和 *te* 因子引入到特征权重计算中并提出两个新的特征权重计算方案：*tf.icf.te* 和 *tf.rf.icf.te* 。

a) *tf.icf.te* 方案的计算公式如下：

$$tf.icf.te(t_i,d_j,c_k) = tf(t_i,d_j)\times lb(1+\frac{|C|}{cf(t_i)})\times te(t_i,c_k)$$

与 *tf.idf* 相比，*tf.icf.te* 是一种混合的特征权重模型。因为，*tf.idf* 中的两个因子都是在文档这个级别上进行估算。而在 *tf.icf.te* 中，*tf* 因子是在文档级别上，*icf* 因子是在类别级别上进行计算的，*te* 因子是衡量特征在类内的分布情况。

) *tf.rf.icf.te* 方案的计算方法如下：

$$tf.rf.icf.te(t_i,d_j,c_k) = tf(t_i,d_j)\times lb(2+\frac{a}{\max(1,c)})\times lb(1+\frac{|C|}{cf(t_i)})\times te(t_i,c_k)$$

其中：a 为在正类中出现特征 t_i 的文本数量，c 为在负类中出现特征 t_i 的文本数量。可以看出 *tf.rf.icf.te* 方案包含四个因子，*tf* 是原始特征频，*rf* 因子则衡量特征 t_i 在正相关类别和负相关类别之间的

分布情况，而 *icf* 因子是衡量特征 t_i 在各类别之间的分布情况，*te* 因子是衡量特征在类内的分布情况。

本文从维吾尔文文本数据集中选择四个特征词 “مەيدان”（场）、“مۇسابقە”（比赛）、“دارۋاز”（达瓦孜）以及 “مۇقام”（木卡姆），分别采用四种权重计算方案计算四个特征词的权重，以考察比较四种权重计算方案的效果。前两个特征词是跟 “体育” 类有关，而后两个特征词是跟 “文艺” 类有关。表 1 和 2 分别显示的是不同两个类别上的特征词权重计算结果。

表 1　五种权重计算方案在 “体育” 类上的计算结果

	tf.idf	*tf.icf*	*tf.rf*	*tf.icf.te*	*tf.rf.icf.te*
مەيدان(4)	1.027	2.023	2.616	0.614	0.771
مۇسابقە (4)	0.992	1.994	2.257	0.709	0.793
دارۋاز(1)	1.723	3.107	1.217	0.023	0.044
مۇقام (2)	1.204	1.635	1.314	0.048	0.055

表 2　五种权重计算方案在 “文艺” 类上的计算结果

	tf.idf	*tf.icf*	*tf.rf*	*tf.icf.te*	*tf.rf.icf.te*
مەيدان(4)	1.027	2.023	1.048	0.216	0.114
مۇسابقە (4)	0.992	1.994	1.016	0.036	0.096
دارۋاز(1)	1.723	3.107	4.107	0.521	0.667
مۇقام (2)	1.204	1.635	3.886	0.443	0.527

在表格中，每个特征词后面括号里的数字代表该特征词的 *cf* 值。从表 1 和表 2 中可以看到，当用 *tf.idf* 和 *tf.icf* 来计算权重时，四个特征词在两个不同类别上的权重值是一样的。这是因为这两个特征权重方案不考虑正相关和负相关类别，而只考虑特征的全局特性（*idf* 和 *icf*）。当用 *tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 来计算权重时，这三种方案能够在两个类别中正确地区分四个特征词。值得注意的是，特征词 “دارۋاز” 的类别频率为 1，也就是说，该特征词只有在“文艺”这一类中出现。当分别用 *tf.icf.te* 和 *tf.rf.icf.te* 计算权重时，在 “文艺” 类中，特征词 “دارۋاز” 的权重比“体育”类中的权重值分别从 0.023 和 0.044 增大到了 0.521 和 0.667，增幅较大。

3　实验及分析

3.1　数据集

本文实验选用维吾尔文文本数据集 Ucorp_A，Ucorp_A 是平衡数据集（Balanced Dataset），包括政治、经济、体育、旅游、教育、文艺、法制、农业、医药保健和计算机等 10 个类别，每个类别有 300 篇文本，2/3 用于训练，1/3 用于测试。

3.2　评价标准

分类性能常用的评价指标有准确率(precision)、召回率(recall)和 F1 值等。

P(准确率)=分类正确的文本数/实际分类的文本数

R(召回率)=分类正确的文本数/应有的文本数

F1=2PR/(P+R)

本文实验使用常用的分类性能评价标准微平均 F_1 值（MicroF_1 值）。

3.3　实验结果

首先，采用 *tf.idf*、*tf.rf* 和 *tf.icf.te*、*tf.rf.icf.te* 四种特征加权方法对数据集中的文本特征进行加权，并且对加权后的样本空间分布进行详细的比较。然后，通过对使用 *tf.idf*、*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 四种特征权重计算方法的 Naive Bayes、kNN、Centroid 和 SVM 四个常用分类器的分类性能比较分析，评价所用四种特征权重计算方法的效果。

实验 1　采用 *tf.idf*、*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 权重计算方案加权的样本空间分布比较实验

不同的权重计算方法，对数据在样本空间中的分布有着不同的影响。也就是说，加权方式改变了样本空间的分布。为了更好地考察本文提出的权重计算方案能不能有效改善样本空间的分布状态以及更好地跟已有的权重计算方法进行比较，本文在维吾尔文文本数据集 Ucorp_A 上，采用传统的 *tf.idf* 方法和较新的 *tf.rf* 方法与 *tf.icf.te* 和 *tf.rf.icf.te* 权重计算方法进行比较实验。

对维吾尔文文本数据集 Ucorp_A，按如下步骤进行实验：

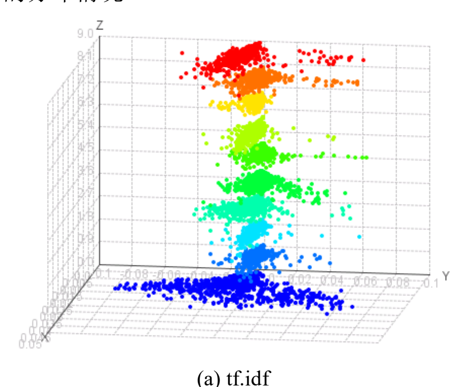
a)对数据集进行预处理，首先，识别单词、删除停用词及非维吾尔文字符、对字符数低于 3 和高于 24 的维吾尔文单词进行过滤、过滤掉 *tf* 值小于 3 的单词，然后，对剩下的单词进行词干提取，最后形成原始特征项集合；

b)采用文献[14]提出的基于类别分布差异和特征熵（class distribution difference and term entropy，CDDTE）的特征选择方法对原始他征集进行特征选择，形成最佳特征子集；

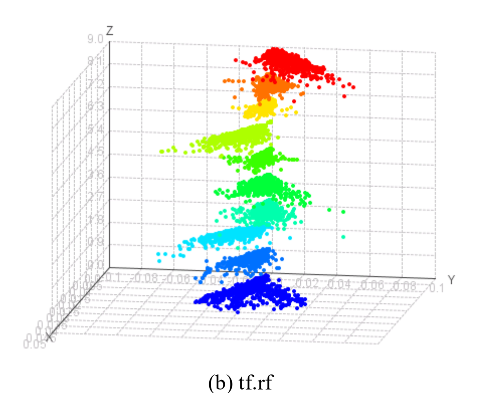
c)分别采用不同的四种特征权重计算方法 *tf.idf*、*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 对特征子集中的每个特征进行权重计算，得到特征子集的四种不同的向量表示形式。

d)特征子集进行向量化表示之后，将高位的特征空间映射到较低维的空间中，并且在较低维数的特征空间中对样本分布进行计算并可视化分析。

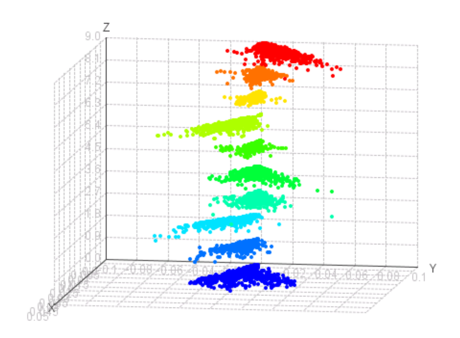
实验结果如图 1 所示。在图 1 中，显示了数据集 Ucorp_A 上分别采用四种特征权重计算方案对特征项赋权后的样本在特征空间中的分布情况。



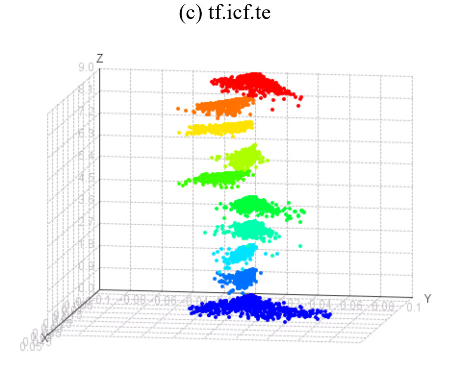
(a) tf.idf



(b) tf.rf



(c) tf.icf.te



(d) tf.rf.icf.te

图 1　采用四种权重计算方案的 Ucorp_A 样本空间分布

从图 1 中可以明显的看出，对于数据集 Ucorp_A 中的样本，四种特征权重计算方法都可以改善样本的空间分布状态，使同类样本更加紧凑，异类样本更加松散，从而简化从样本到类别的映射关系。其中，*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 特征权重计算方法的效果都要好于 *tf.idf* 方法，且 *tf.rf.icf.te* 特征权重计算方法的效果明显好于 *tf.rf* 加权方法的效果。

实验 2 采用 tf.idf、tf.rf、tf.icf.te 和 tf.rf.icf.te 权重计算方案加权的分类效果比较实验

特征权重计算方法的好坏对文本分类效果影响很大，因此可以通过对文本分类效果的比较来评价特征权重计算方法的效果。

本实验仍然使用维吾尔文文本数据集 Ucorp_A。对数据集采用 CDDTE[14]方法进行特征选择，用 *tf.idf* 方法来计算特征权重构造文本向量进行了分类实验。当特征数为 2000 的时候四个分类器的 MicroF_1 值都达到了各自的最高值。在本实验中，我们使用这 2000 个单词作为文本特征，分别采用不同的三种特征权重计算方法 *tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 来对每个特征进行加权，形成三种不同的特征空间。最后在不同的特征空间中对数据集的测试文本进行向量化，分别采用四种分类器 NB、kNN、Centroid 和 SVM 对测试文本进行分类实验。表 4 给出了在 Ucorp_A 上的分类实验结果。

表 4　四种特征权重计算方案在 Ucorp_A 上的分类结果

	tf.idf	*tf.rf*	*tf.icf.te*	*tf.rf.icf.te*
Naive Bayes	0.8363	0.8478	0.8427	0.8507
kNN	0.8110	0.8512	0.8363	0.8733
Centroid	0.8741	0.9031	0.8952	0.9249
SVM	0.9063	0.9153	0.9105	0.9353

从表 4 可以看出，在数据集 Ucorp_A 上，*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 等三种权重方案在所有四个分类器的分类实验中都表现出比 *tf.idf* 好的性能。其中，*tf.rf.icf.te* 方法效果最好。与 *tf.idf* 相比，四个分类器与 *tf.rf.icf.te* 方法相结合的分类 MicroF_1 值都有了明显的提高。其中 kNN 分类器 MicroF_1 值的提高幅度最大，达到了 6.23%，其次是 Centroid 分类器，达到了 5.08%，而 SVM 和 NB 分别为 2.9%和 1.44%。观察发现，*tf.rf* 方案和四种分类器组合的分类效果比 *tf.idf* 和 *tf.icf.te* 要好，但明显低于 *tf.rf.icf.te* 方案。此外，在 kNN 和 Centroid 分类器下，与 *tf.idf* 相比，*tf.icf.te* 加权的分类 MicroF_1 值分别提高了 2.53%和 2.11%。而 *tf.icf.te* 方案在 NB 和 SVM 分类器下的分类 MicroF_1 值提高幅度很小，跟 *tf.idf* 基本持平。原因是因为，SVM 分类器分类效果主要受核函数的影响较大，而特征权重方法的影响小。

通过实验可以发现，与已有的特征权重计算方案相比，本文的两个特征权重计算方案 *tf.icf.te* 和 *tf.rf.icf.te* 能够改善维吾尔文文本分类的效果。

4　结束语

本文提出了引入类别信息和特征熵的权重计算方案 *tf.icf.te* 和 *tf.rf.icf.te*。在维吾尔文文本数据集 Ucorp_A 上，将 *tf.icf.te* 和 *tf.rf.icf.te* 特征加权方法与 *tf.idf* 和 *tf.rf* 两种特征加权方法进行了比较实验。使用 *tf.idf*、*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 四种特征加权方法对数据集中的文本特征计算权重，通过绘制加权后的样本空间分布图，定性地说明了 *tf.icf.te* 和 *tf.rf.icf.te* 加权方法能够使同类样本更加紧凑，异类样本更加松散；该方法能够明显改善样本的空间分布状态，从而简化从样本到类别的映射关系。使用 Naive Bayes、kNN、Centroid 和 SVM 四种常用的分类器，在上述数据集上对 *tf.idf*、*tf.rf*、*tf.icf.te* 和 *tf.rf.icf.te* 四种特征加权方法的效果进行比较实验，实验结果表明，本文的 *tf.icf.te* 和 *tf.rf.icf.te* 权重计算方法能够提高维吾尔文文本分类的微平均 F_1 值。

参考文献：

[1] 张爱华，靖红芳，王斌，等. 文本分类中特征权重因子的作用研究 [J]. 中文信息学报, 2010, 24 (3): 97-104. (Zhang Aihua, Jing Hongfang, Wang Bin, *et al.* Research on effects of term weighting factors for text categorization [J]. Journal of Chinese Information Processing, 2010, 24 (3):

97-104.)

[2] Feng Guozhong, Li Shaoting, Sun Tieli, *et al.* A probabilistic model derived term weighting scheme for text classification [J]. Pattern Recognition Letters, 2018, 110 (1): 23–29.

[3] Chen Kewen, Zhang Zuping, Long Jun, Zhang Hao. Turning from TF-IDF to TF-IGM for term weighting in text classification [J]. Expert Systems With Applications, 2016, 66 (1): 245-260.

[4] 陈科文，张祖平，龙军. 文本分类中基于熵的词权重计算方法研究 [J]. 计算机科学与探索, 2016, 10 (9): 1299-1309. (Chen Kewen, Zhang Zuping, Long Jun, *et al.* Research on entropy-based term weighting methods in text categorization [J]. Journal of Frontiers of Computer Science & Technology, 2016, 10 (9): 1299-1309.)

[5] Fattah M A, Sohrab M G. Combined term weighting scheme using FFNN, GA, MR, Sum, & average for text classification [J]. International Journal of Scientific & Engineering Research, 2016, 7 (8): 2031-2039.

[6] Debole F, Sebastiani F. Supervised term weighting for automated text categorization [C]// Proc of ACM Symposium on Applied Computing. New York: ACM Press, 2003: 784-788.

[7] Pei Zhuli, Shi Xiaohu, Marchese M, *et al.* An enhanced text categorization method based on improved text frequency approach and mutual information algorithm [J]. Progress in Natural Science, 2007, 17 (12): 1494-1500.

[8] Lan Man, Tan C L, Su Jian, *et al.* Supervised and traditional term weighting methods for automatic text categorization [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31 (4): 721-735.

[9] Liu Ying, Han Tongloh, Sun Aixin, *et al.* Imbalanced text classification: a term weighting approach [J]. Expert Systems with Applications, 2009, 36 (1): 690-701.

[10] Quan Xiaojun, Liu Wenyin, Qiu Bite *et al.* Term weighting schemes for question categorization [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33 (5): 1009-1021.

[11] 贾隆嘉，孙铁利，杨凤芹，等. 基于类空间密度的文本分类特征加权算法 [J]. 吉林大学学报, 2017, 35 (1): 92-97. (Jia Longjia, Sun Tieli, Yang Fengqin *et al.* Class space density based weighting scheme for automated text categorization [J]. Journal of Jilin University, 2017, 35 (1): 92-97.)

[12] 张羚，陆余良，杨国正. 基于词频类别相关的特征权重算法 [J]. 计算机应用研究，2017, 34 (2): 386-391. (Zhang Ling, Lu Yuliang, Yang Guozheng, *et al.* Categories-related term weighting method based on term frequency [J]. Application Research of Computers, 2017, 34 (2): 386-391.)

[13] 李学明，李海瑞，薛亮, *et al.* 基于信息增益与信息熵的 TFIDF 算法 [J]. 计算机工程, 2012, 38 (8): 37-40. (Li Xueming, Li Hairui, Xue Liang, *et al.* TFIDF algorithm based on information gain and information entropy [J]. Computer Engineering, 2012, 38 (8): 37-40.)

[14] 阿力木江·艾沙，吐尔根·依布拉音，库尔班·吾布力等. 基于类别分布差异和特征熵的维吾尔语文本特征选择 [J]. 计算机应用研究, 2013, 30 (10): 2958-2961. (Alimjan Aysa, Turgun Ibrahim, Kurban Ubul, *et al.* Uyghur text feature selection based on class distribution difference and term entropy [J]. Application Research of Computers, 2013, 30 (10): 2958-2961.)